(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 685 835 A1**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **95107651.2**

(22) Date of filing: **19.05.95**

(51) Int. Cl.^{6}: **G10L 5/06**, G10L 7/08, G10L 9/06, G10L 9/18

(30) Priority: **30.05.94 FI 942518**

(43) Date of publication of application:
06.12.95 Bulletin 95/49

(84) Designated Contracting States:
AT BE CH DE DK ES FR GB GR IE IT LI NL PT SE

(71) Applicant: **TECNOMEN OY**
Finnoonniitynkuja 4
SF-02270 Espoo (FI)

(72) Inventor: **Ranta, Jari**
Niittylaakso 3
FIN-02760 Espoo (FI)

(74) Representative: **Pfister, Helmut, Dipl.-Ing.**
Buxacher Strasse 9
D-87700 Memmingen (DE)

(54) Speech recognition based on HMMs.

(57) A speech recognition method that combines HMMs and vector quantization to model the speech signal and adds spectral derivative information in the speech parameters is presented. Each state of a HMM is modelled by two different VQ-codebooks. One is trained by using the spectral parameters and the second is trained by using the spectral derivative parameters.

Rank Xerox (UK) Business Services
(3.10/3.09/3.3.4)

Problem that is to be solved

The accuracy of speaker independent speech recognition is inadequate with current algorithms especially when the recognition is done through dialled-up telephone lines. The accuracy of a speech recogniser means the ability to recognise an utterance by comparing it to the system's precomputed word templates.

Current solutions and their drawbacks

Traditionally Hidden Markov Models (HMM) that are based on probability theory are used in speech recognisers. During the recognition phase a probability that a certain model can produce the utterance is computed. The model that has the highest probability is selected as the recognised word.
Reference [2] presents a speech recognition method that uses vector quantization (VQ) with HMMs instead of statistical pattern matching. During the recognition phase the squared error is computed between the word template and given utterance. Word templates are HMMs where each state has its own VQ-codebook. Every VQ-codebook is computed from the training data with LBG-vector quantization algorithm [5] and it contains the typical speech parameters that occur in that state. A template that gives the smallest square error is chosen as the recognised word. The modified Viterbi-algorithm that is used in computing the distance is also presented in [2].
A speech recogniser that uses HMMs with continuous mixture densities is presented in [3]. It uses the cepstrum derived from LPC-analysis and its derivative as the speech parameters (spectral derivative). The vector that is computed from speech contains short-term information about spectral changes in the signal (via the cepstrum) and the short-time spectral derivative contains information from longer time span (via the delta cepstrum). By adding the spectral derivative to the speech parameters a more accurate, 2-dimensional presentation of the time-varying speech signal is obtained (frequency and time). According to [3] this enhances the recognition accuracy of HMM-model that uses continuous mixture densities.

However, the recognition accuracy with both of these methods is inadequate.

A patented algorithm [4] that is used for speaker verification gives a 1% false recognition and false rejection rate when using numbers from zero to nine to perform verification. (The reference does not mention how many numbers the user has to speak during the verification process.)

Invention

The idea is to combine the methods presented in [2] and [3], i.e. to add the spectral derivative shown in [3] to the speech parameters of the HMM-VQ algorithm in [2] and to obtain better recognition accuracy.

Speech recogniser in [2] that uses VQ-distortion measure is compared to the known statistical HMMs which use continuous or discrete mixture densities and the superiority is of the HMM-VQ over traditional HMM is clearly shown. Because the use of spectral derivative in statistical HMMs improves recognition accuracy, by adding the spectral derivative to the HMM-VQ model the recognition accuracy can improved even more. The spectral derivatives contain information from longer time period. By combining the two parameter sets a higher recognition rate can be obtained than using the speech spectrum alone as in [2]. During the training process separate VQ-codebooks are computed for speech spectrum and spectral derivatives.

When using test data that was obtained through dialled-up telephone lines the recognition accuracy was higher when compared to method of [2]. A 100 speakers were used for training and 25 different speakers were used for testing. The vocabulary consisted from eleven Finnish words and an accuracy of 98.85 % was obtained. If a threshold was set so that the relative distance between the best and the second best word must be greater than 10% before a valid word is recognised, a 100% accuracy was obtained while 1.5% of the input was rejected. Reference gives a recognition accuracy of 100%, but the test data is recorded over high quality microphone. Speech is much harder to recognise from the telephone because the bandwidth is limited and the frequency responses of the telephone lines can greatly vary.

Example of an application using the invention

The proposed method can be used for speech recognition in the same way as statistical HMMs. The units of speech that are to be recognised can be either words, phonemes, triphones etc.
The application can be for example a voicemail system where the menu commands are given by speech instead of touch tones ("...if you want to listen a message press 3..."). The system is trained to recognise a

small vocabulary of command words which is compared against the speech uttered by the user.

The same algorithm can also be used for speaker verification and preliminary experiments gave promising results. By using just one word for verification the error rate was 0.21 %. The error is computed by multiplying the number of false rejections and the number of false recognitions and taking the square root of the result [6]. Five different 'real' speaker's was used in the test which were compared to 185 impostors. The word models were computed from five repetitions of a word. (Compare this to [4] that gives less than 1 % of false rejection rate). The error rate obtained in [6] was 3.2% and high quality speech samples were used for testing.

The ability to recognise the speaker through telephone is important in voicemail applications when the telephone cannot send DTMF-tones. In such a case there is no other reliable method to recognise the caller than his own voice.

Operational description

A method for discrete speaker independent speech recognition is presented in the following. The recognition method uses HMMs with vector quantization for representing the speech parameters. The HMM is a simple state machine where transition can occur only to current or to the next state. Block diagram of the different phases in speech recognition is given in figure 1. The new speech recogniser uses HMM models so that each state is described by two different VQ-codebooks. One is obtained using the spectral parameters computed with the PLP-analysis [1] and the other is obtained by using the spectral derivative parameters.

Different phases of the recognition

1. Speech analysis

Speech that is to be recognised is analysed with PLP-analysis [1] in 30 ms parts and by using 15 ms intervals. Analysis phase gives speech parameters $cc_l(m)$ where $1 \leq m \leq 5$, representing this 15 ms frame. Vector $cc_l(m)$ at time instant l is weighted with window $W_c(m)$, that results in

$$c_l(m) = cc_l(m)^* W_c(m). \tag{1}$$

2. Computing of the parameters

Reference [3] describes how the use of spectral derivative enhances the recognition accuracy of a statistical HMM. Spectral derivative means the weighted average of spectral parameters obtained from the analysis phase. The average is computed over a short time window according to equation (2)

$$\Delta c_l(m) = \left[\sum_{k=-K}^{K} k c_{l-k}(m) \right] * G, \quad 1 \leq m \leq 5, \ K = 2 \tag{2}$$

where G is an amplification factor selected so that the variances of the vectors $c_l(m)$ and $\Delta c_l(m)$ are equal. The value used here was 0.2. By combining these two vectors a parameter set that describes time frame l is obtained

$$O_l = \{c_l(m), \Delta c_l(m)\} \tag{3}$$

which consists of ten elements. The speech parameter set is denoted by **C** and the spectral derivative parameter is denoted by Δ**C** i.e.

$$\mathbf{C} = \{c_l(m)\} \text{ and } \Delta\mathbf{C} = \{\Delta c_l(m)\} \tag{4}$$

3. Training phase

The word templates are trained separately for spectral parameters **C** and for spectral derivative parameters Δ**C**. Templates are trained by using a vector quantization algorithm and the training process is illustrated in figure 2.

1. Speech samples that are used for training are first analysed with PLP-analysis and the vectors **C** and Δ**C** are obtained. These vectors describe the speech at 15 ms intervals. Each analysed speech sample is first divided linearly into states so that each state have equal amount of vectors. The states correspond to states in a HMM. If a HMM with 8 states is needed, each analysed speech sample is divided in eight parts of equal length. This linear division gives a starting point for the training algorithm.
2. A separate codebook is computed for each state in a HMM. The vector quantization algorithm is applied on every vector on every state from every sample. For example, all the vectors that belong to state one in every speech sample are used to create the codebook for state one. The same is done for states from two to eight. The codebooks contain a set of vectors that give the minimum square distance between the vectors used for training. There are several algorithms to design a vector quantizer, a method presented in [5] is used here.
3. When the optimal codebooks are computed from training vectors, the VQ-distortion of each speech sample from the model is computed. The sum denotes the 'goodness' of the model. The smaller the total distortion, the better the model represents the words that were used when the model was created.
4. The sum is compared to the sum obtained from the previous iteration. If the sum is larger than the previous sum, training ends.
5. If the new sum is smaller, the speech samples are divided into new set of states and the learning process continues from step 2. The optimum state sequence is found by using the Viterbi-algorithm.

It is important that the speech samples are collected from the same environment where the recogniser is intended to be used. If there is a need to recognise speech through telephone, then the word templates must be trained with words that are collected through telephone. If different environments are used in training and in recognising, recognition accuracy will degrade substantially.

4. Distance calculation

Distance of a word is computed between the uttered speech and word template. Distance D from each word template is computed with the modified Viterbi algorithm [2] according to following equation.

$$D = \min_{x} \sum_{i=1}^{L} \left\{ d\left(c_i, VQ_{x_i}\right) + d\left(\Delta c_i, VQ_{x_i}\right) + d\left(x_{i-1}, x_i\right) \right\}. \qquad (5)$$

Here $x_0 x_1 \ldots x_L$ means the state sequence of a HMM and

$$VQ_{x_i}$$

means codebook at state x_i. The number of states in HMM is denoted by S i.e. $1 \leqq x_i \leqq S$. L is the number of frames in a word. $d(x_{i-1},x_i)$ is zero if there is a transition from state x_{i-1} to state x_i, otherwise $d(x_{i-1},x_i) = \infty$.

$$d\left(c_i, VQ_{x_i}\right)$$

denotes the smallest distance between vectors c_i and

$$VQ_{x_i}$$

which is defined as

$$d\left(c_i, VQ_{x_i}\right) = \min_j d\left(c_i, vq_j\right), vq_j \in VQ_{x_i}. \qquad (6)$$

$d(c_i, vq_j)$ is defined as

$$d\left(c_i, vq_j\right) = \sum_{m=1}^{5} \left(c_i(m) - vq_j(m)\right)^2 \qquad (7)$$

Here vq_j is a component in the codebook. The number of components in one codebook

$$VQ_{x_i}$$

is a power of two, usually 16 or 32. Equation (7) computes the smallest square error between the vector c_i and the codebook component vq_j. Equation (6) computes the smallest square error between the vector c_i and codebook

$$VQ_{x_i}.$$

Similarly

$$d\left(\Delta c_i, VQ_{x_i}\right)$$

denotes the smallest distance between vectors Δc_i and

$$VQ_{x_i}.$$

The Viterbi algorithm given in (5) is computed recursively so that the VQ-distortion is added at the same time for both parameter sets. I.e. the smallest square distance computed for the spectral parameters and for the spectral derivative parameters. This can be written as

$$\begin{aligned} &g(j,t) = \min_i \left\{ g(i,t-1) + d\left(c_t, VQ_j\right) + d\left(\Delta c_t, VQ_j\right) + d(i,j) \right\} \\ &D = g(S,L), \quad t = 1,2,\ldots,L \ \text{ ja } \ j = 1,2,\ldots,S. \end{aligned} \qquad (8)$$

Here is the idea of the invention in mathematical form. The term $d(\Delta c_i, VQ_j)$ is added to the Viterbi algorithm and equation (8) is computed for each 15 ms time frame t. There is also a possibility to use two HMMs where the first is computed using the spectral parameters and the second is computed using the spectral derivative parameters. The results from these two models are then added together with appropriate weighting to obtain the final recognition result.

References:

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", Journal of Acoustical Society of America, Vol. 87, No. 4, April 1990, pp. 1738-1752.

[2] S. Nakagawa and H. Suzuki, "A new speech recognition method based on VQ-distortion measure and HMM", ICASSP-93, pp. II-676 - II-679.

[3] L. R. Rabiner, J. G. Wilpon and F. K. Soong, "High performance connected digit recognition using hidden Markov models", IEEE Transactions on Acoustics Speech and Signal Processing, vol. 37, pp. 1214-1225, August 1989.

[4] High Accuracy Speaker Verification System, Ensigma Ltd

[5] Y. Linde, A. Buzo, R. M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, vol. COM-28, NO. 1, January 1980.

[6] D.K.Burton, "Text-Dependent Speaker Verification Using Vector Quantization Source Coding", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-35, No. 2, pp. 133-143, February 1987.

Claims

1. A speech recognition method based on Hidden Markov Models (HMM) that uses PLP-analysis for computing the speech parameters, wherein the VQ-distortion of the speech parameters from the codebooks is computed and the codebooks are generated by using the LBG-algorithm, the word template that gives the smallest VQ-distortion being selected as the recognised word, **characterized** in that the LBG-algorithm is used to train the word models separately for spectral parameters and spectral derivative parameters and that separate codebooks are used for both parameter sets which in turn are used to model each state of a HMM.

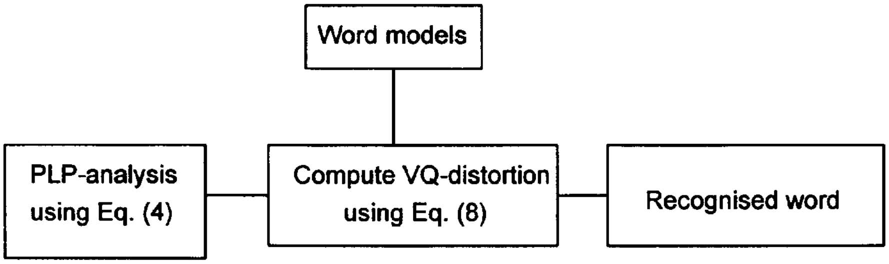
Word models
PLP-analysis using Eq. (4)
Compute VQ-distortion using Eq. (8)
Recognised word

Figure 1. Block diagram of the different phases of speech recognition.

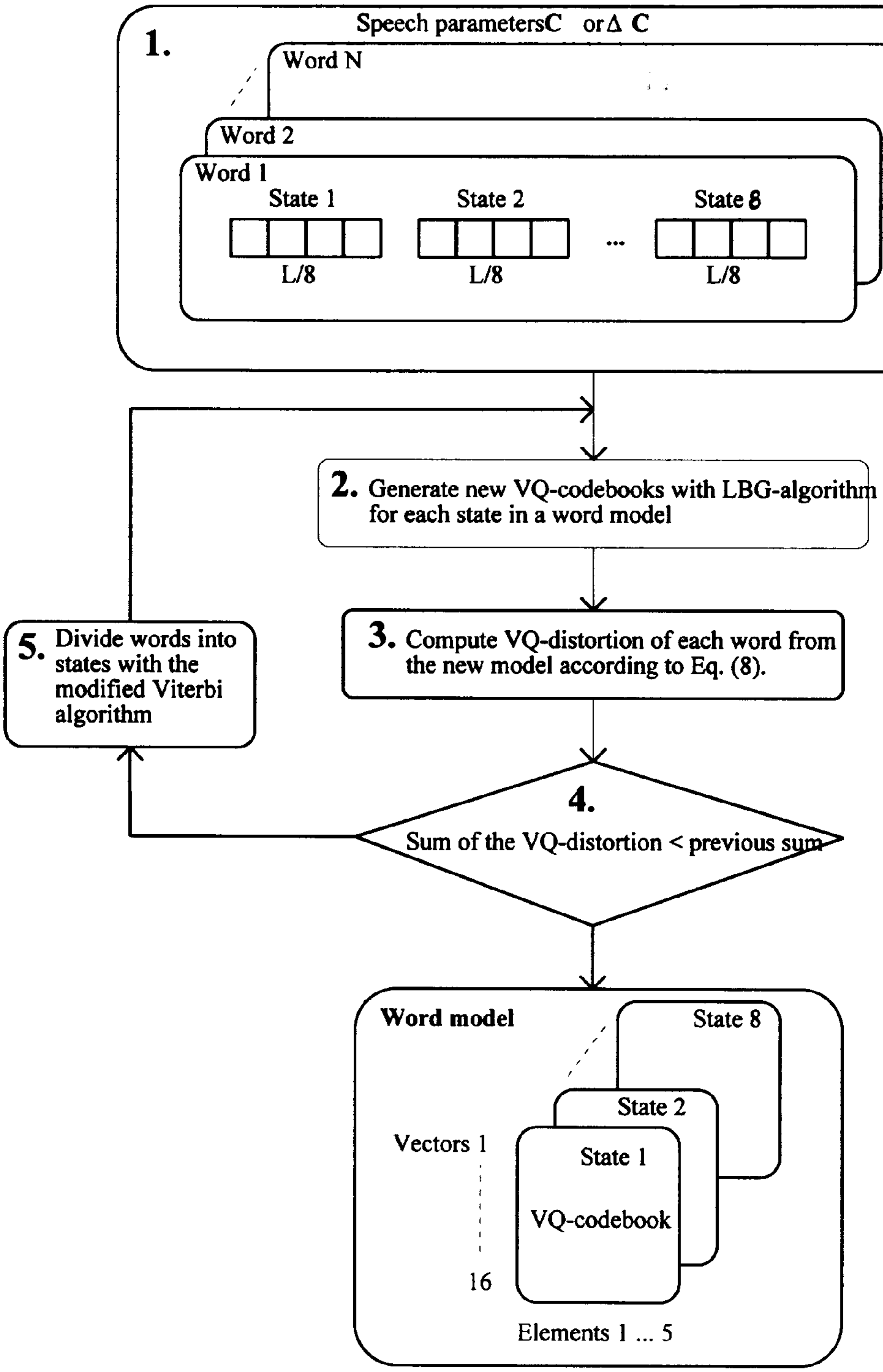
1.
Speech parametersC or Δ C
Word N
Word 2
Word 1
State 1
State 2
State 8
...
L/8
L/8
L/8
2. Generate new VQ-codebooks with LBG-algorithm for each state in a word model
3. Compute VQ-distortion of each word from the new model according to Eq. (8).
4. Sum of the VQ-distortion < previous sum
5. Divide words into states with the modified Viterbi algorithm
Word model
State 8
State 2
State 1
VQ-codebook
Vectors 1
16
Elements 1 ... 5

Figure 2. Block diagram of the training process. Both parameter sets are trained separately.

European Patent Office

EUROPEAN SEARCH REPORT

Application Number

EP 95107651.2

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl. 6)
A	EP - A - 0 560 378 (TOSHIBA K.K.) * Fig. 3; abstract; claim 1 * --	1	G 10 L 5/06 G 10 L 7/08 G 10 L 9/06 G 10 L 9/18
A	EP - A - 0 562 138 (IBM) * Fig. 1; abstract; claim 1 * --	1	
A	EP - A - 0 590 925 (IBM) * Fig. 1; abstract; claim 1 * ----	1	
			TECHNICAL FIELDS SEARCHED (Int. Cl.6)
			G 10 L 3/00 G 10 L 5/00 G 10 L 7/00 G 10 L 9/00
	The present search report has been drawn up for all claims		

Place of search	Date of completion of the search	Examiner
VIENNA	16-08-1995	BERGER

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document